Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP10/2011

To Smooth or Not to Smooth?
The Case of Discrete Variables in
Nonparametric Regressions

L´eopold Simar & Valentin Zelenyuk

Date:
December 2011

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

To Smooth or Not to Smooth? The Case of Discrete Variables in Nonparametric Regressions

Léopold Simar* Valentin Zelenyuk**

December 16, 2011

Abstract

In a seminal paper, Racine and Li, (*Journal of Econometrics*, 2004) introduce a tool which admits discrete and categorical variables as regressors in nonparametric regressions. The method is similar to the smoothing techniques for continuous regressors but uses discrete kernels. In the literature, it is generally admitted that it is always better to smooth the discrete variables. In this paper we investigate the potential problem linked to the bandwidths selection for the continuous variable due to the presence of the discrete variables. We find that in some cases, the performance of the resulting regression estimates may be deteriorated by smoothing the discrete variables in the way addressed so far in the literature, and that a fully separate estimation (without any smoothing of the discrete variable) may provide significantly better results, and we explain why this may happen. The problem being posed, we then suggest how to use the Racine and Li approach to overcome these difficulties and to provide estimates with better performances. We investigate through some simulated data sets and by more extensive Monte-Carlo experiments the performances of all the proposed approaches and we find that, as expected, our suggested approach has the best performances. We also briefly illustrate the consequences of these issues on the estimation of the derivatives of the regression. Finally, we exemplify the phenomenon with an empirical illustration. Our main objective is to warn the practitioners of the potential problems posed by smoothing discrete variables by using the so far available softwares and to suggest a safer approach to implement the procedure.

Keywords: Discrete regressors, Nonparametric regression, Kernel smoothing, Cross-validation, Local Polynomial Estimator

JEL Classification: C1,C14, C13

*Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium; email leopold.simar@uclouvain.be. Financial support from the "Inter-university Attraction Pole", Phase VI (No. P6/03) of the Belgian Government (Belgian Science Policy) and from the Center for Efficiency and Productivity Analysis (CEPA), School of Economics, The University of Queensland, Australia are gratefully acknowledged.

**School of Economics and Centre for Efficiency and Productivity Analysis, The University of Queensland, Australia; email v.zelenyuk@uq.edu.au.

1 Introduction

Rapid advance of computing power and wider availability of large data sets encouraged many researchers to substantially increase their attention to various non-parametric methods for estimating regression relationships. One of the most popular of such non-parametric methods appears to be the local polynomial least squares method, considered by Stone (1977), Cleveland (1979), Cleveland and Delvin (1988), Fan (1992, 1993), Fan and Gijbels (1992), Ruppert and Wand (1994) and popularized by Fan and Gijbels (1996). This method received even greater appeal when it was substantially empowered by the seminal work of Racine and Li (2004), who suggested a neat way to deal with discrete regressors in the context of nonparametric regression.[1] This work inspired many interesting applications in a wide range of areas, for example, by Stengos and Zacharias (2006), Maasoumi et al. (2007), Parmeter et al. (2007), Eren and Henderson (2008), Walls (2009), Hartarska et al. (2010), Henderson (2010), to mention just a few. In these and other works that used Racine and Li (2004) approach, researchers were able to obtain new insights with much more confidence, as their approach was free from imposing any parametric form on the regression relationship, while using both continuous and discrete regressors without splitting the sample into sub-samples for each value of the discrete variables. In all the empirical as well as theoretical studies and software codes that use Racine-Li approach that we are aware of, the way this approach is applied is in its "default" or simple form that we will describe below.

Indeed, it became somewhat common to smooth the discrete regressors in local polynomial least-squares almost automatically, perceiving that one should obtain better results than if the estimator is applied to each group, identified by the discrete variable, separately (see Li and Racine, 2007 and the references therein). For instance, in Racine and Li (2004, p.113), one can read:

> *"One should expect that the smoothing method outperforms the frequency method in general, since the former includes the latter as a special case (when $\lambda = 0$). However, when the sample size is very large, the computational cost can be high for the cross-validation-based smoothing method. Therefore, in practice one may want to use the frequency method when the sample size is much larger than the number of discrete cells due to the computational simplicity of the frequency method. But even in such a situation the efficiency gain of the smoothing method over the frequency method can be substantial because the cross-validation method may choose large values of λ for some discrete variables (e.g., Insik et al., 2002)."*

While this statement appears to hold in various cases and is supported by all simulations of Racine and Li (2004) for the case of local constant fitting, in this article we illustrate that in some cases it may not be the case and smoothing the discrete variables can actually

[1] They extended the Aitchison and Aitken (1976) ideas for smoothing discrete variables and provided all the asymptotic theory. Other basic references on smoothing discrete variables are Titterington (1980) and Wang and Van Ryzin (1981).

deteriorate substantially the resulting estimator of the regression function. In some situations even, a fully separate estimation for each group identified by the discrete (categorical) variable may give much more accurate results (e.g., in terms of Mean Squared Error, MSE) than the approach with smoothing over the discrete variable. In these cases, the reduction in variance, or the efficiency gain due to smoothing of the discrete regressors, can be well outweighed by a substantial bias introduced due to this smoothing. This may happen both for small as well as for relatively large samples, and so, for such cases, it may be preferable to make a fully separate estimation for each group.

We will see below that the source of the problem comes from the bandwidth structure suggested by the basic or "default" method appearing in all papers in the literature we are aware of (see some references above) and the various softwares implementing the Racine-Li approach. In this basic approach, a "simple" bandwidth scheme is proposed for the continuous variables, in the sense that the same bandwidth vector is taken across the various subgroups determined by the discrete variables and then, the bandwidths for the continuous and for the discrete variables are simultaneously determined. So that, even if the resulting estimator of the bandwidths for a discrete variable takes a value zero (i.e., implying no smoothing of this discrete variable with separate estimation by groups), the resulting bandwidths for the continuous variables are still restricted to be common to the various categories of this discrete variable. This may lead in some cases to "over-smoothing" in some groups and "under-smoothing" in the others.

To fix the ideas, suppose that we use local constant kernel method (Nadaraya-Watson) and that we have two groups of observations (determined by one discrete variable) and only one continuous variable. Suppose in addition that in one group, the variable is relevant (derivatives of the regression w.r.t. this variable are not zero) and in the other it is not relevant. In the later group, the optimal bandwidth would converge to infinity (see e.g., Hall et al., 2007), whereas in the first group, it has to converge to zero. A fully separate analysis ("no smoothing" the discrete variable) will capture this feature whereas the "simple smoothing", suggested by the basic implementation of Racine and Li technique that appears in all applications we are aware of, will miss this feature. The latter approach will provide a common bandwidth for the continuous variable that will under-smooth the regression in the group where the variable is irrelevant and over-smooth the regression in the group where it is relevant. This extreme case seems obvious but apparently it has been overlooked in practice. We can also imagine less extreme situations where the phenomenon would be similar: small influence of the continuous variable in one group and more complex structure in the other.

Now, using local polynomial smoothing, the same problem may appear if in one group the local polynomial approximation is not far from the true regression, but the structure in the

other is globally much more complex (e.g., for local linear case, if the true regression is linear, the optimal bandwidth has to converge to infinity, see Li and Racine, 2004). Obviously, the problem can even be more severe when estimating the derivatives of the regression. We will briefly illustrate this in one example below, showing the bad consequences the simple-smoothing could lead to. To the best of our knowledge, this phenomenon has never been analyzed in the literature using smoothing techniques for discrete variables. It is one of the objective of our paper to investigate the empirical consequences of these issues.

There is a simple way to improve the method and overcome these potential difficulties. One can allow, in the bandwidth selection procedure, different bandwidth parameters for the continuous variables in each category of the discrete variables. We will call this method the "complete-smoothing" approach to contrast it to the default "simple-smoothing" used so far in the literature and to the "no-smoothing" approach where the groups are treated fully separately. This richer approach is of course at a cost of computational complexity, but we will see below that the gain in precision of the regression estimate can be substantial. We will limit the presentation to the case of categorical discrete variables, so each value of the discrete variables determines a group category.

More generally and beyond the extreme cases described above, whether the bias beats the variance, or not, essentially depends on the degree of difference of curvatures of the regression relationship pertinent to each group identified by the discrete variable and to some extent also depends on other aspects of the DGP (Data Generating Process), such as the size of the noise, etc. Thus, in general, *a priori* it is not clear whether it is better to smooth the discrete variables or to do a fully separate estimation (unless the latter is hardly reliable or impossible due to very small data in a given group) and, so far, there appears to be no formal rule of thumb for deciding on this dilemma. The complete-smoothing approach, that we suggest below, allows for smoothing the discrete variables but also uses different bandwidths for each group for the continuous variable. Since this encompasses both the no-smoothing approach (fully separate estimation procedure) and the simple-smoothing technique as special cases, we can expect theoretically better performances of the complete-smoothing approach.

Because *a priori* it is not clear whether all categories of a discrete variable must have a common bandwidth, our finding is very important for practitioners, as it warns against an automatic use of smoothing over the discrete regressors in a simple way (as provided by the available softwares), without considering that this might actually produce less accurate estimation results. More importantly, we hope our paper will stimulate further research in this area to find some theoretically justified ways (e.g., statistical tests or rules of thumb) for deciding whether "to smooth or not to smooth" over the discrete regressors and whether or not to go with the more general, but more computationally demanding, method we suggest

in this paper.

Our paper is organized as follows: Section 2 recalls the basic notations and results of Local Linear Least-Squares (LLLS) methods which we use to illustrate the issue;[2] Section 3 focuses on the potential source of problems when smoothing the discrete variables and suggests an extension that would, in theory, provide better performances; Section 4 illustrates how severe can be the problem by some simple visualized examples and by some more extensive Monte-Carlo experiments and how the complete-smoothing approach outperforms the other approaches; Section 5 illustrates the issue with a real data set and finally, Section 6 concludes and summarizes our main findings.

2 Description of the Method

The point we stress in this paper is a general phenomenon linked to nonparametric regression, but we illustrate it on a method that appears to be the most popular in practice, the local linear least squares (which is a particular case of Local Polynomial Least Squares, LPLS). We summarize here the basics of the method and we refer to textbooks for details (e.g., see Li and Racine, 2007, Fan and Gijbels, 1996 or Pagan and Ullah, 1999). The idea of local polynomial smoothing is to allow flexible form for approximating locally the true unknown regression. Formally, we assume that the dependent endogenous variables $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$, are generated according to the following regression model:

$$Y_i = m(X_i) + \varepsilon_i, \tag{2.1}$$

where $X_i = (Z_i, Z_i^d)$, with $Z_i \in \mathbb{R}^p$ being continuous and Z_i^d being a L-dimensional discrete variable. We will focus on the presentation for categorical unordered variables, but the same could be done for naturally ordered variables by using appropriate kernels, see Racine and Li (2004) for details. In addition, we make the standard assumptions on the errors, ε_i, that they are independent random variables with $E(\varepsilon_i \mid X_i) = 0$, $E(\varepsilon_i X_i) = 0$ and $V(\varepsilon_i \mid X_i) < \infty$, although the phenomenon we will discuss may apply to more sophisticated setups.

The flexibility of the model is related to the fact that the unknown regression function $m(\cdot)$ is not specified. No particular assumptions are made on m itself except for some smoothness properties on $m(\cdot, z^d)$. For the sake of simplicity, we will assume that $m(\cdot, z^d)$ is twice continuously differentiable in its first p continuous arguments. Finally, we need also

[2] Our remarks and suggestions could obviously be applied also when using higher order local polynomial smoothing or other non-parametric kernel-based regression methods, such as local-non-linear least squares (see Gozalo and Linton, 2000) or local maximum likelihood techniques with discrete variables as in Park et al. (2010), etc.

some regularity (smoothness) of the density $f_{Z|Z^d}(\cdot; z^d)$ with respect to the p continuous arguments.

The main idea of LPLS is to approximate $m(u, v)$ for all (u, v) in a neighborhood of a given point (z, z^d) by a local polynomial of degree r in the direction of z and then select the parameters of this local polynomial by minimizing the resulting sum of the squared errors. A degree $r = 0$ would provide the local constant (Nadaraya-Watson type) estimator. For simplicity of notation, we will limit our presentation to the case of local linear approximations ($r = 1$). Extension to higher orders follows the same ideas but at a cost of notational complexity. So we take the following local approximation:

$$m(u, v) \approx \alpha_{z,z^d} + \beta'_{z,z^d}(u - z), \tag{2.2}$$

where $\alpha_{z,z^d} \in \mathbb{R}$ and $\beta_{z,z^d} \in \mathbb{R}^p$ are quantities to be estimated that, in general, vary with (z, z^d). To take only neighboring observations around (z, z^d), or to give more weights to them, when evaluating the least-squares criterion, the kernel approach is used. For the continuous variables we use a product kernel (but many other multivariate kernels would also work), i.e.,

$$K_h(Z_i - z) = \prod_{j=1}^{p} \frac{1}{h_j} K_j\left(\frac{Z_{i,(j)} - z_{(j)}}{h_j}\right), \tag{2.3}$$

where $h = (h_1, \ldots, h_p)$ is a vector of bandwidths, $z_{(j)}$ is the j^{th} component of z and $K_j(\cdot)$ is a standard univariate kernel function (e.g., univariate standard Gaussian density). For the discrete variables we use the Racine and Li (2004) kernel, i.e.,

$$\Lambda_\lambda(Z_i^d, z^d) = \prod_{\ell=1}^{L} \lambda_\ell^{\mathbb{1}\left(Z_{i,(\ell)}^d \neq z_{(\ell)}^d\right)}, \tag{2.4}$$

where $\mathbb{1}(A)$ is the indicator function, with $\mathbb{1}(A) = 1$ if A holds, and 0 otherwise, and $\lambda_\ell \in [0, 1]$ are bandwidths for the discrete variables $\ell = 1, \ldots, L$. The local least squares criterion at a given point (z, z^d) measuring the quality of the approximation is thus given by

$$C_n(\alpha_{z,z^d}, \beta_{z,z^d}; z, z^d) = \sum_{i=1}^{n} \left(Y_i - (\alpha_{z,z^d} + \beta'_{z,z^d}(Z_i - z))\right)^2 K_h(Z_i - z)\Lambda_\lambda(Z_i^d, z^d), \tag{2.5}$$

We note that if for a particular ℓ, we have $\lambda_\ell = 0$ (with the convention that $0^0 = 1$), then there is no smoothing of this ℓ^{th} discrete variable; i.e., the evaluation in (2.5) is done separately for each subsample determined by this discrete variable, with a common h. At the other limit, if $\lambda_\ell = 1$, we do not take into account this discrete variable in the analysis; i.e., all the sample points have weight in (2.5) independent from the value of $Z_{i,(\ell)}^d$.

Let $\widehat{\alpha}_{z,z^d}$ and $\widehat{\beta}_{z,z^d}$ minimize the criterion C_n at (z, z^d), then the proposed estimator of the regression function at the point (z, z^d), denoted by $\widehat{m}(z, z^d)$, is given by $\widehat{\alpha}_{z,z^d}$ whereas $\widehat{\beta}_{z,z^d}$ gives an estimate of the first partial derivatives of m with respect to the continuous variables z evaluated at (z, z^d).

The selection of appropriate bandwidths (h, λ) can be done by Least-Squares Cross Validation (LSCV) method, although many other approaches can be adopted (e.g., corrected AIC method as also considered by Li and Racine, 2004, etc.). When adopting the LSCV approach, the values $\hat{h}$ and $\hat{\lambda}$ are the values that minimize

$$CV(h, \lambda) = \sum_{i=1}^{n} \left(Y_i - \widehat{m}_{(-i)}(Z_i, Z_i^d)\right)^2 M(Z_i, Z_i^d), \tag{2.6}$$

where $M(Z_i, Z_i^d)$ is a weight function trimming out boundary observations and $\widehat{m}_{(-i)}(Z_i, Z_i^d)$ is the leave-one-out kernel estimator of $m(Z_i, Z_i^d)$, i.e., estimated by using (2.5), but leaving the i^{th} observation out of the sample. The properties of the final resulting estimator $\widehat{m}(z, z^d)$ when using these bandwidths are described in Racine and Li (2004) (see Theorem 2.3) for the local constant case ($r = 0$) and in Li and Racine (2004) for the local linear approximations (see Theorem 2.2). It is important to notice that, as common in all these smoothing approaches, these theorems assume that $h_j \to 0$ for all $j = 1, \ldots, p$, and $nh_1 \ldots h_p \to \infty$ as $n \to \infty$. As pointed out above, if $\hat{\lambda}_\ell = 0$ for all $\ell = 1, \ldots, L$ we have $\Lambda_\lambda(Z_i^d, z^d) = \mathbb{1}(Z_i^d = z^d)$, so that the estimation in (2.5) is done only with the data having this value z^d, so the estimation of the regression function is done separately for each group, with common h.

The argument generally admitted in the literature so far is that it is always better to smooth the discrete variable in (2.5), because the separate analysis (called the frequency method by Racine and Li, 2004) on each separate subsample defined by the categorial variables Z^d would correspond to the particular case when all $\lambda_\ell = 0$, $j = 1, \ldots, L$ (see e.g., the quote in the introduction above). We indicate in the next section that it might not be the case in all situations, including very simple ones.

3 Over- and Under-Smoothing Problem and A Simple Solution

To simplify the argument, let us suppose that we have only one discrete variable defining two groups of observations (the argument would be the same when considering all the subgroups defined by the $L \geq 1$ categorical variables Z^d). Suppose that in addition, the DGP in the two groups have different characteristics (shape of the regression function, or curvature, or

size of the noise, etc.). Extreme cases of such differences have been shortly described in the introduction. Unless the sample size within one of the group is very small, the bandwidth selection procedure described above may provide small values of $\hat{\lambda}$ (for relevant discrete regressor), resulting in the frequency method in the limiting case where $\hat{\lambda} = 0$ (two separate samples), and a common h for different groups.

As pointed out above for the extreme cases, this could be inappropriate in many situations where some important characteristics of the DGP are quite different in the two subgroups. In such instances, having common bandwidths $\hat{h}$ for different groups identified by z^d, may force under-smoothing (over continuous variables) for one group while over-smoothing for the other. While this may not matter asymptotically (as long as the common $\hat{h}$ has the proper order), it happens to matter in finite samples (small and even relatively large ones), sometimes substantially, as illustrated in our examples in the next section. So, in this case, it might be better to do fully separate estimation within each subgroup, allowing the vector of bandwidths for continuous variables to vary across the different groups (no smoothing at all for the discrete variables). As pointed out by Racine and Li (2004), this may increase the variance but this will lower the bias; at the end, it could provide estimators with smaller MSE, than the case with smoothing over the discrete variable, even if the smoother-selection procedure leads to $\hat{\lambda} \approx 0$. We will show in the next section that the loss in accuracy of this default "simple-smoothed" estimator may be dramatic. Of course, if the sample size in one group is too small one cannot hope to get sensible results with a separate nonparametric estimation. So, a solution is needed.

A natural way to overcome these problems of over- and under-smoothing in subgroups is to allow smoothing over the discrete variables, as in Racine and Li (2004), but to proceed to what we call a complete-smoothing, i.e., to allow also different bandwidths for the continuous variables in the two groups. Formally, in the case of two groups defined by one discrete variable, the equation (2.5) defining the estimator could be replaced by

$$C_n(\alpha_{z,z^d}, \beta_{z,z^d}; z, z^d) = \sum_{i=1}^{n} \big(Y_i - (\alpha_{z,z^d} + \beta'_{z,z^d}(Z_i - z))\big)^2 \\ \times \left[K_{h_1}(Z_i - z)\mathbb{1}\big(Z_i^d = z^d(1)\big) + K_{h_2}(Z_i - z)\mathbb{1}\big(Z_i^d = z^d(2)\big)\right] \Lambda_\lambda(Z_i^d, z^d), \quad (3.1)$$

where $z^d(k)$, $k = 1, 2$ are the two values of z^d defining the two subgroups (e.g. $z^d(1) = 1$, $z^d(2) = 0$). For simplicity, we keep the same kernel function in the two groups, but we allow potentially different bandwidths h_1 and h_2 for these groups. Optimal bandwidths $(\hat{h}_1, \hat{h}_2, \hat{\lambda})$ could be derived in a similar way as above in (2.6) but at a computational cost if p is large and if product kernels are used for the continuous variables or if L or the number of groups identified by the discrete variables is large.

It must be clear that the general formulation we propose in (3.1) encompasses both the fully separate analysis by groups (this is, in our complete-smoothing setup, the case when $\lambda = 0$), and the simple-smoothing approach (this is, imposing $h_1 = h_2$ in our complete-smoothing setup). So, in theory, we could only improve the performances of the resulting estimator relative to these two particular cases. Note also that the mixture of two kernel functions we use in (3.1) shares (under the same regularity conditions) the same properties as the kernel used for continuous variables in Racine and Li (2004) and Li and Racine (2004). So, all the asymptotic results derived there remain valid, as long as the bandwidths h_1 and h_2 have the appropriate sizes when both sample sizes in each group go to infinity.

In the next section we will illustrate and visualize the problem discussed here in some simple examples. We will investigate the finite sample properties of the different approaches. This will confirm the expected theoretical performance of our extension over the traditional smoothing approach and over the fully separate approach. We find that the gain of precision may be substantial in practice.

4 Illustration

We first present some very simple examples that allow us vividly illustrating the issue raised in the preceding section. We will provide "typical" pictures resulting from particular simulated samples (generated according to the scenarii described below). Of course we cannot conclude general statements about one simulated sample but the idea is to provide visualization of the problem. Afterwards, we will confirm what we see by a more detailed Monte-Carlo experiment. In all presented simulations, we considered the following regression relationship (although we tried many others and obtained the same conclusions)

$$Y_i = a_1 + a_2 Z_i^d + b_1 Z_i + b_2 Z_i^d Z_i + b_3 Z_i^2 + b_4 Z_i^d Z_i^2 + b_5 Z_i^d \sin(\pi Z_i) + \varepsilon_i. \qquad (4.1)$$

Varying the choice of the parameters will provide the various examples explored in this section.[3]

[3]It is worth to mention that in the simulations we used the Aitchison and Aitken (1976) discrete kernel, which is equivalent to the Racine and Li kernel used in Section 2 (see for details, Racine and Li, 2004). Here the discrete smoothing parameter λ takes its values in $[0, 0.5]$ in place of $[0, 1]$. Note also that since the range of Z is $[-2, 2]$, we limit the search of optimal bandwidths of Z (h, h_1 and h_2) in the range $(0, 20]$. This does not change the picture and the MC results. In all the cases, the bandwidths h, h_1 and h_2 have been scaled by s_Z, the empirical standard deviation of Z. So the bandwidths h, h_1 and h_2 used while optimizing (2.5) and (3.1) are the values reported in the Figures and the Tables below, multiplied by s_Z.

Example 1: linear vs. periodic regression

For the first example presented, we set $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0, b_4 = 0, b_5 = 2$, with $\varepsilon_i \sim N(0, \sigma_{\varepsilon,i}$, where $\sigma_{\varepsilon,i} = 2 - Z_i^d$. Here, for each simulation, the $Z_i \sim U(-2, 2)$ for the continuous variable Z and the discrete variable Z^d was set randomly at 1 if $W > 0.25$ and set at 0 if $W \leq 0.25$, where $W \sim U(0, 1)$. So, we randomly obtained about 75% of observations for group 1 (with $Z_i^d = 1$) and about 25% for group 2 (with $Z_i^d = 0$).[4]

In this first example, in group 2 ($z^d = 0$) we have a linear model with more noisy data and smaller sample size and for group 1 ($z^d = 1$), we have a linear model (slightly different intercept and slope) plus a cyclic component. In Figure 1, we present a typical result of the estimation by using the 3 approaches described above: Approach 1 is a fully separate estimation of the two subsamples (no-smoothing of the discrete variable), Approach 2 is the simple-smoothing, i.e., smoothing the discrete variable, with common bandwidth for the continuous regressor Z and Approach 3 is the complete-smoothing, i.e., smoothing the discrete variable but keeping potentially different bandwidths for the continuous variable. Figure 1 presents typical results just for one sample of a moderate size $n = 100$ and one sample with a large size $n = 400$ (similar pictures have been obtained for other sizes, see the Monte-Carlo experiment below).

Looking first to the left panels ($n = 100$), we can see that the simple-smoothing (Approach 2) suffers from a serious drawback in this scenario and that the no-smoothing estimation (Approach 1) gives much better results both for $n = 100$ and $n = 400$. The complete-smoothing of Approach 3, encompassing the 2 preceding ones, does as well as the fully separate analysis for these samples. The fully separate estimation substantially outperforms the estimation with simple-smoothing over the discrete regressor, as the latter approach under-smoothes for the group 1 and slightly over-smoothes for the group 2. Note that for this example, the under-smoothing is more pronounced because the group 1 dominates in the pooled sample by its larger size and so the common bandwidth selected in the CV optimization for (h, λ) is relatively close to what is optimal for the group 1 in the separate estimation, while for the group 2, the true optimal bandwidth must in fact go to infinity to attain the correctly specified parametric model. Of course the complete-smoothing (Approach 3), allowing different bandwidths for the continuous regressor Z in the two groups, corrects for this.

One may argue that this is a small sample problem, in fact it is not. We see clearly in the left panel of Figure 5 that when $n = 1000$, the final estimator of the regression lines by using Approach 2 still behaves poorly in group 2, due to the persistent under-smoothing. Of

[4]Most conclusions remain the same for other compositions (e.g., 50% vs. 50%, etc.).

course, with such large samples, allowing different bandwidths for the continuous variable in the two groups gives close to a perfect fit, both with Approach 1 and Approach 3. We cannot say the same about Approach 2 even for larger samples.

Results of the Monte-Carlo experiment summarized in Table 1 confirms what has been seen for two particular samples illustrated in Figure 1. Allowing n going from 50 to 400, we did 200 Monte-Carlo (MC) replications in each case. The table provides $\overline{AMSE}$, the mean of the Approximate Mean Squared Error ($AMSE$) of each sample, obtained over the 200 MC replications. For one MC sample, the $AMSE$ is defined as

$$AMSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{m}(Z_i, Z_i^d)\right)^2.$$

The table also gives the estimated standard deviation of the Monte-Carlo estimate of the mean, defined as

$$std_{MC} = \frac{1}{\sqrt{M}}\sqrt{\frac{1}{M-1}\sum_{k=1}^{M}\left(AMSE_k - \overline{AMSE}\right)^2},$$

where $\overline{AMSE} = (1/M)\sum_{k=1}^{M} AMSE_k$ and M is the number of MC replications. This std_{MC} allows to check if the differences observed in the table for the $\overline{AMSE}$ are significant.

We remark that all the figures appearing in Table 1 vary as expected when n increases. It is worth noting that in all the simulations for this scenario, the CV yielded $\hat{\lambda}$ that is very close to zero for both Approach 2 and Approach 3, and that Approach 3 gives systematically less weight to the smoothing of the discrete variable than Approach 2.

We can see in Table 1 that over many replications, the overall $\overline{AMSE}$ is always (significantly) smaller for the Approach 1 than for Approach 2 (often twice smaller), and that this does not vanish when the sample size increases. Note that the difference in $\overline{AMSE}$ is much larger for the smaller group, where as also is seen from the figures and as explained above, the problem of under-smoothing is more sever due to domination of the larger group in the CV selection of the common bandwidth of the traditional Racine-Li approach. The $\overline{AMSE}$ for Group 2 is significantly smaller for Approach 1 and 3 than for Approach 2, and this is true even for small samples, as small as $n = 50$ (with on the average, $n_2 \approx 12$) and the difference being about three times for $n = 100, 200, 400$. Looking to the median of the bandwidth selected by the different approaches, we see clearly that the Racine-Li Approach 2 under-smoothes the data of Group 2. We see also that complete-smoothing Approach 3 is, as expected and explained above, very safe here since it gives the same results as Approach 1 (no significant differences here). Note also that, as a consequence of the theory provided by Racine and Li (2004) and Li and Racine (2004), in all the cases, the $AMSE$ reduces

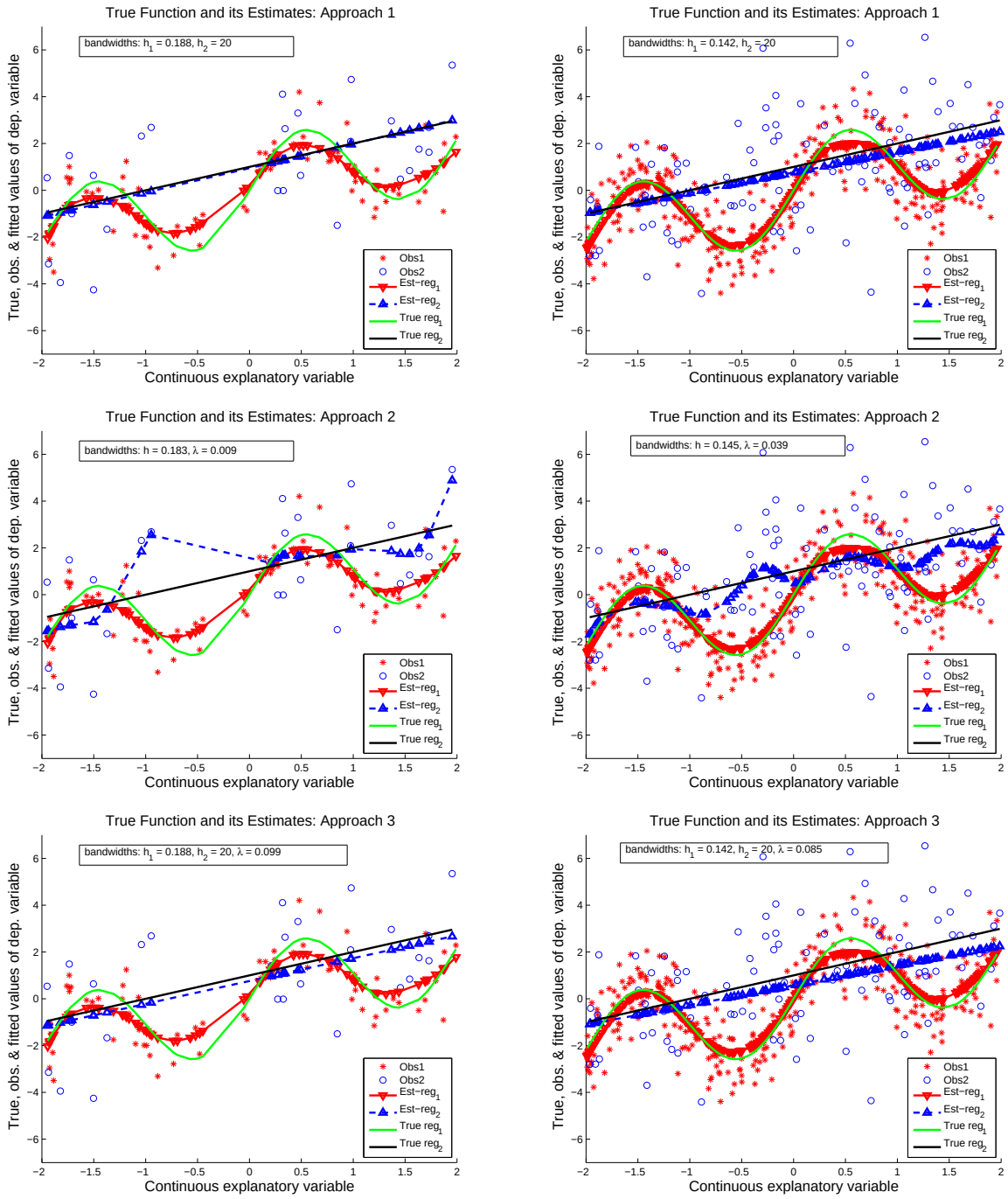

Figure 1: *Example 1: Left panel,* $n = 100$ *and right panel,* $n = 400$*. From top to bottom: Approach1 (no-smoothing the discrete variable), Approach2 (simple- smoothing), Approach3 (complete-smoothing).*

as n increases and that the optimal bandwidths (except $\hat{h}_2$ when computed separately) go to zero as n goes to infinity. Again, the LSCV procedure for Approach 3 gives very sim-

ilar bandwidths for h_1 and h_2 to those obtained by Approach 1, while the estimate of λ is persistently smaller than for the Approach 2, indicating that Approach 2 suggests more similarities between groups than suggested by Approach 3.

Example 2: linear vs. quadratic regression

It may look like the phenomenon we notice is pertinent only to cases with radically different curvatures. So, for the second example we take much more similar regression lines for the two groups. In equation (4.1) we selected the values $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0, b_4 = 1, b_5 = 0$, all the other elements of the scenario of the first examples are the same. So here, in group 1 ($Z_i^d = 1$) we have replaced the periodic part by a quadratic term, implying some curvature. The scenario for group 2 ($Z_i^d = 0$) remains the same (i.e., linear) as in the first example.

Figure 2 below shows typical examples with the resulting fits of the 3 approaches, with $n = 100$ and $n = 400$. The pictures tell us the same story as in Example 1, although the difference in curvature is not so radical. This is also confirmed by the Monte-Carlo experiments (200 MC replications) and the results are displayed in Table 2. To summarize and to save place, we can say that most of the comments coming from Example 1 can be replicated. Here again, Approach 1 gives always better results than the simple-smoothing Approach 2. The complete-smoothing Approach 3 is much safer here as well: it is always better than Approach 2 and in most of the cases, is even significantly better than Approach 1, for all sample sizes. Here, the CV-estimated values of λ are larger for Approach 3 than Approach 2, indicating that leaving more flexibility to the choice of the bandwidth for the continuous variable in the two groups increases the gain of precision while smoothing the discrete variable. This is because the two regression lines are not so different as they were in Example 1.

Example 3: linear vs. very similar quadratic regression

We shortly describe another scenario where the quadratic regression is quite similar to the linear one: we diminish the size of the shift (intercept) from -1 to -0.5 and the coefficient of quadratic term is decreased from 1 to 0.25. So, in equation (4.1) we selected the values $a_1 = 1, a_2 = -0.5, b_1 = 1, b_2 = 0.1, b_3 = 0, b_4 = 0.25, b_5 = 0$, all the other elements of the scenario are the same. We do not present the pictures because the two regression lines are so close that we do not learn too much by looking to a particular sample realization. However, the Monte-Carlo presented in Table 3 is not without interest.

Globally, and as expected here, the simple-smoothing technique (Approach 2) behaves

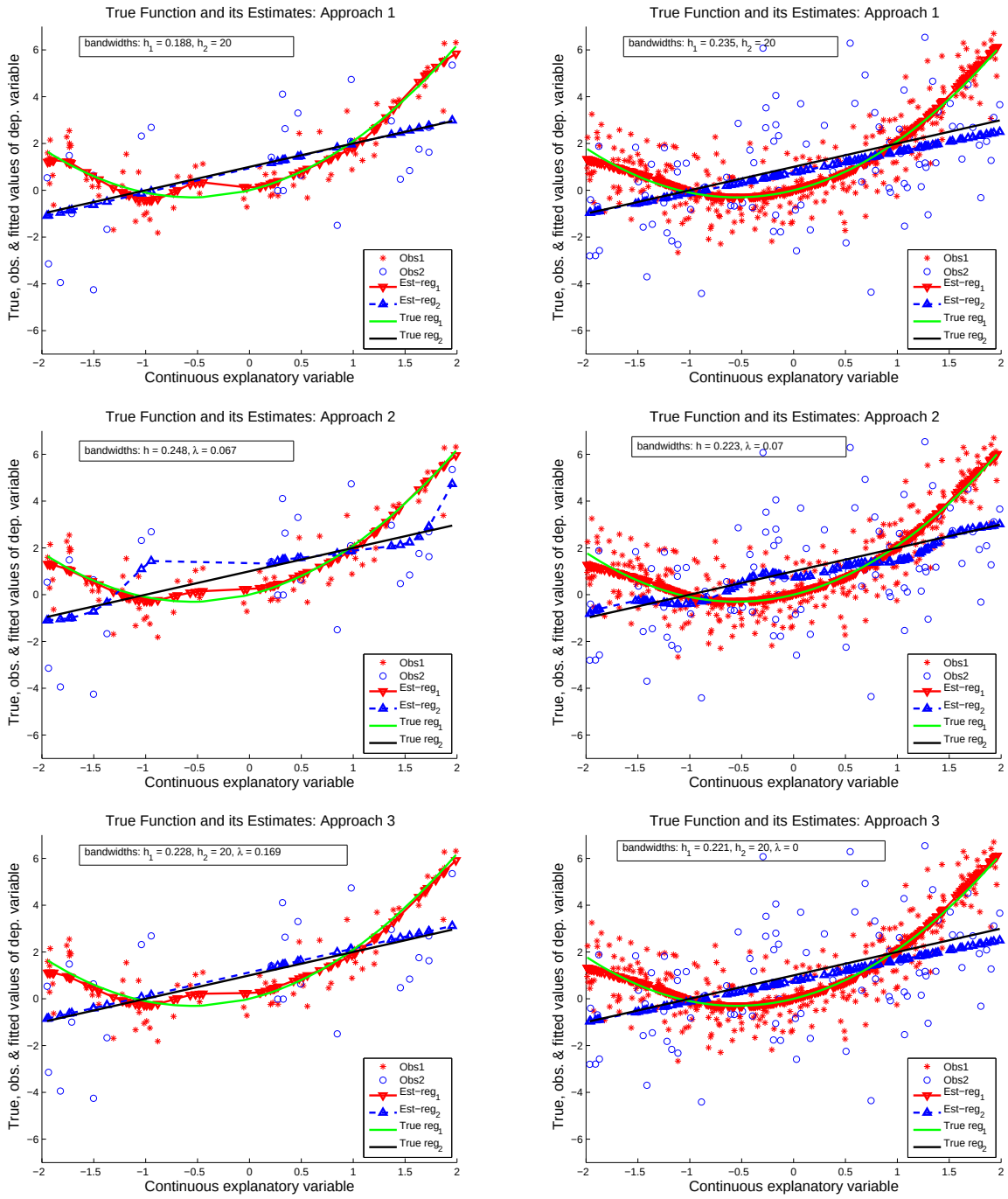

Figure 2: *Example 2: Left panel,* $n = 100$ *and right panel,* $n = 400$*. From top to bottom: Approach1 (no-smoothing the discrete variable), Approach2 (simple-smoothing), Approach3 (complete-smoothing).*

barely significantly better: for small sample size ($n = 50$) it outperforms the no-smoothing case (Approach 1), mostly due to the gain of precision in estimating the regression relation-

ship for the group 2. The gain of variance is larger than the loss due to bias. However, the dominance of Approach 2 over Approach 1 vanishes very quickly as n increases (no more significant differences from $n = 100$). But the most interesting result is that even in this scenario, favorable to Approach 2, the results obtained by using the complete-smoothing technique (Approach 3) always suggest significantly better performance for the two other approaches. This is true even for $n = 50$ (mostly for better estimation in group 2) and this does not vanishes when n increases.

Example 4: quadratic vs. quadratic regression

In this scenario, both regression relationships are quadratic and so, the theoretical optimal values of the bandwidths in both groups should converge to zero when the sample size increases. We keep a difference in the shift between the two regressions as in Example 1 and 2, a slight difference of the linear component but one regression has a quadratic component which is two times larger than the first. Specifically we have in equation (4.1) the values $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0.15, b_4 = 0.15, b_5 = 0$, all the other elements of the scenario of the first examples are the same. So we expect in this example to have a better behavior of the simple-smoothing technique (Approach 2) than Approach 1, because the curvatures of the two regressions are quite similar.

Figure 3 below shows typical examples with the resulting fits of the 3 approaches, with $n = 100$ and $n = 400$. We see indeed that Approach 2 behaves relatively well compared to the two others, although, the group 2 seems again to be under-smoothed for the same reasons explained above for Example 1 and 2. Looking to the right panel of Figure 4, we see that even for larger samples ($n = 1000$), Approach 2 does not provide estimators so close to the true regressions, as Approaches 1 and 3 do. The disappointing behavior of the estimators resulting from Approach 2 becomes clearer when looking to the Monte-Carlo experiment for this scenario. Table 4 indicates that Approach 2 does as well as Approach 1 when the total $\overline{AMSE}$ is considered, but at a cost of balancing a better behavior for group 2 and a worse behavior for group 1. Here again, Approach 3 is certainly the safest, in all the cases it shows similar or better performances than the two other approaches.

Example 5: quadratic vs. periodic regression

Finally, we tried a quadratic vs. a periodic regression in the two groups: this is similar to Example 1, except that now the linear model is wrong in both groups. We select in equation (4.1) the values $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0.25, b_4 = 0, b_5 = 2$, all the other elements of the scenario of the first examples are the same. Typical samples and

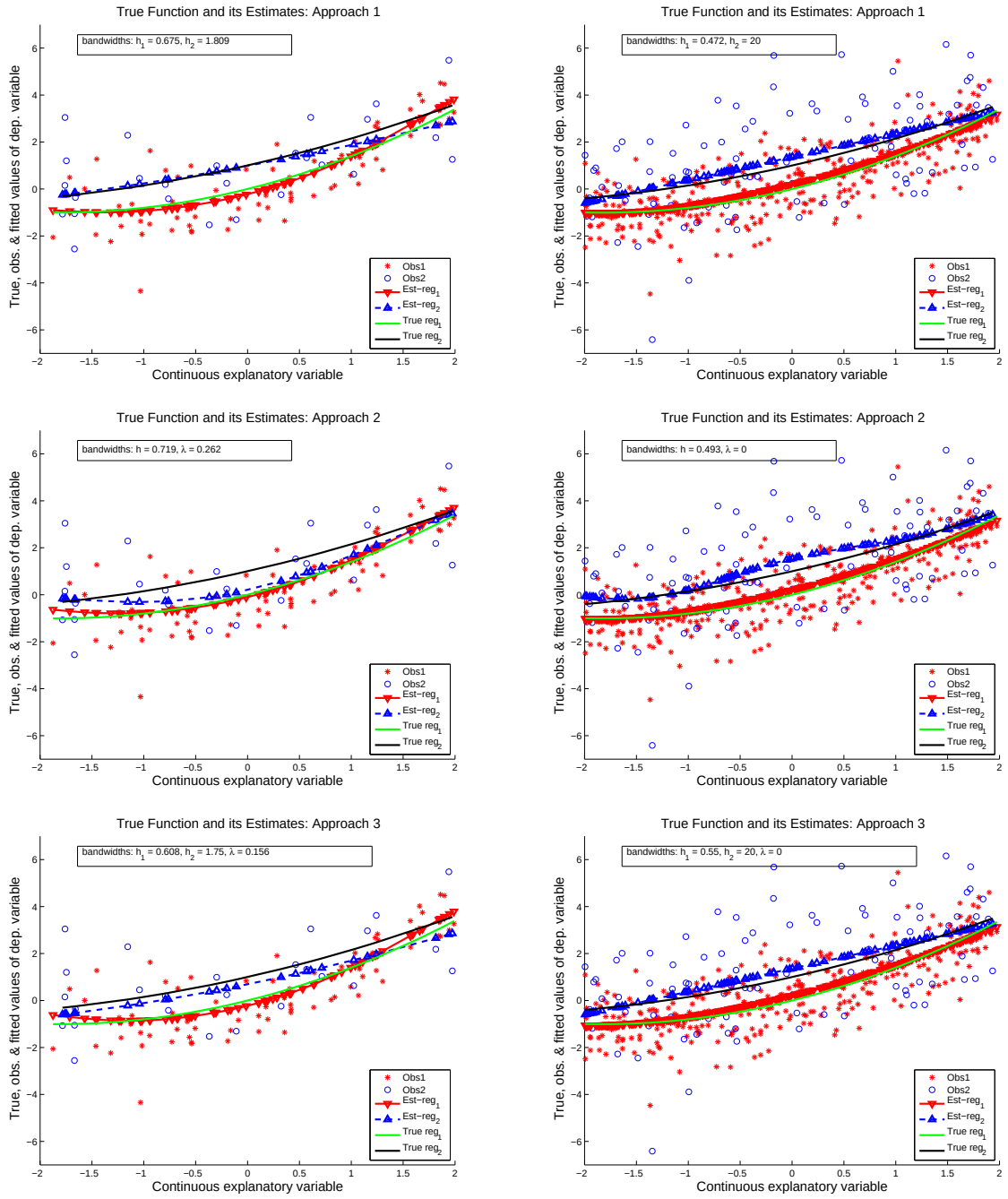

Figure 3: *Example 4: Left panel,* $n = 100$ *and right panel,* $n = 400$*. From top to bottom: Approach1 (no-smoothing the discrete variable), Approach2 (simple- smoothing), Approach3 (complete-smoothing).*

results of estimation are displayed in Figure 5 and the full picture is given in the Monte-Carlo Table 5. We see indeed that in this case the difference between Approach 2 and the

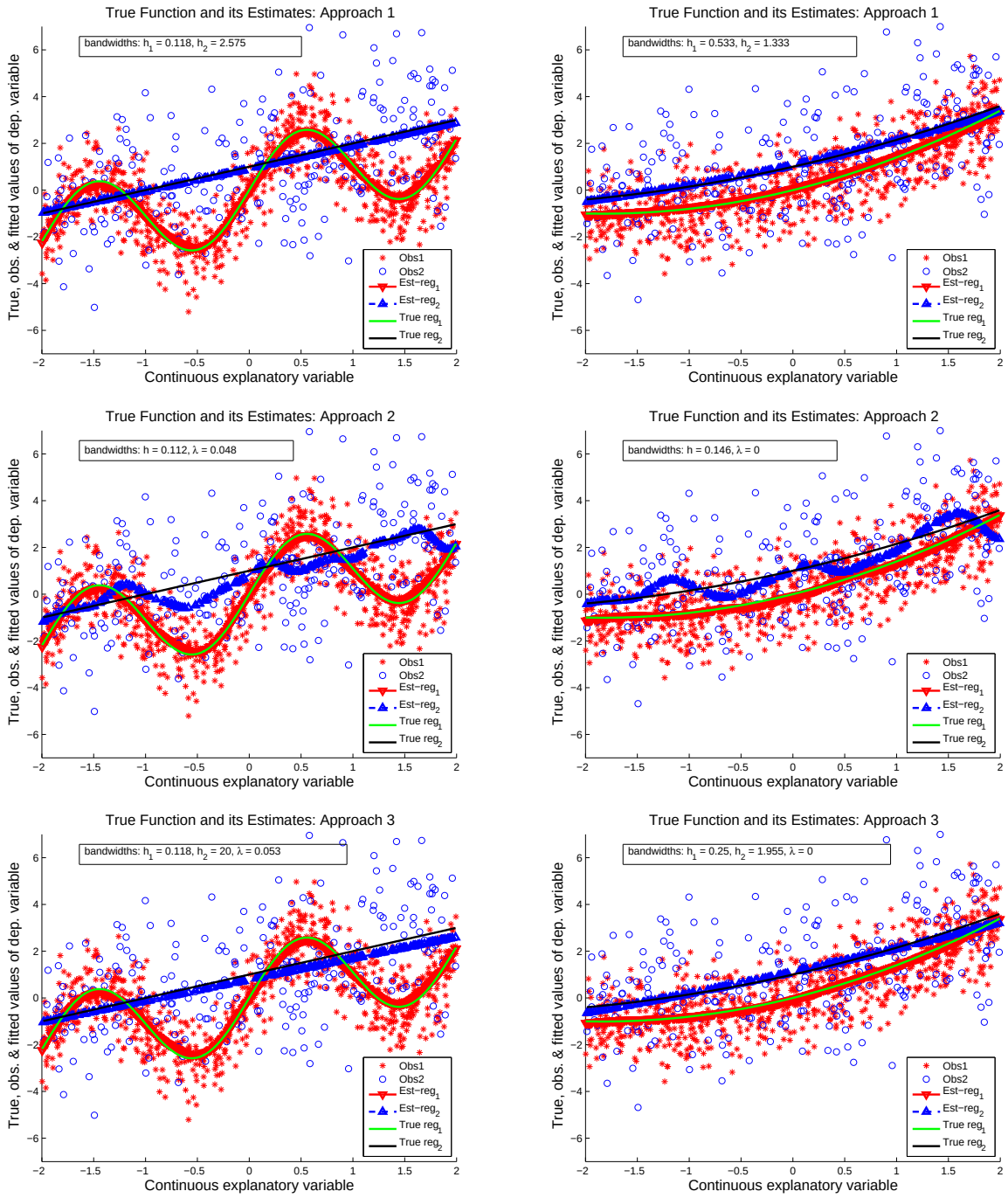

Figure 4: *Left panel, Example 1 with* $n = 1000$ *and right panel, Example 4 with* $n = 1000$. *From top to bottom: Approach1 (no-smoothing the discrete variable), Approach2 (simple-smoothing), Approach3 (complete-smoothing).*

two other approaches is not so dramatic as in the previous examples (because here, in both groups, the optimal bandwidth for the continuous component should converge to zero when

n increases). Yet, looking to the MC results in Table 5, we confirm the general comments given for Example 1. Approach 1 has always better performance than Approach 2 and again, Approach 3 is the safest way to approach the estimation in this scenario (with significantly better performances).

Consequences on the estimation of derivatives

Obviously, the problem of derivatives estimation is contaminated by the phenomenon we describe. To save space we only illustrate this in the case of the scenario described in Example 1. Figure 6 displays one typical sample and the resulting estimates of the first partial derivatives using the 3 approaches. This figure illustrates clearly that the estimation of derivatives and the related estimates of marginal effects, elasticities, etc. can be even more severely flawed by using the simple-smoothing approach. Indeed, as one can clearly see from Figure 6, with simple smoothing one obtains radically varying and even changing the sign estimates of the derivatives for the group where their true values are constant or vice versa. The problem sustains whether the total sample size is 100 or 400 (or more). This means that research conclusions, policy implications and, consequently, the real policy decisions based on such estimates can be misleading, wrong and perhaps even damaging. Note that for this same example, the complete-smoothing approach produced much better results, very close to the true values. We also did a more complete Monte-Carlo experiment that confirmed largely these expected results, but to save space we omit them from the paper.

Concluding remarks from the MC experiments

While we presented only particular examples here, there are of course many others, some of which we also tried, where the same phenomenon is observed. We presented here some cases, like Example 1, where one relationship is periodic while the other is linear just for vividly illustrating the point, showing how substantial could be the difference in performance. But we also observed that the same phenomenon may still be present when the degree of non-linearity is not very different across groups, i.e., linear vs. quadratic, as well as when both are quadratic but with different curvatures or when one is periodic and the other is quadratic. In our simulations, we also observed that the problem of over and under-smoothing in different groups reduces when the difference in non-linearity or non-smoothness in regression relationship across groups reduces but such information can hardly be known in practice. In other words, the "default" Racine-Li approach (simple-smoothing) implicitly restricts the groups to have similar "degree of smoothness", which is less restrictive than parametric assumptions, yet it is more restrictive than fully non-parametric estimation.

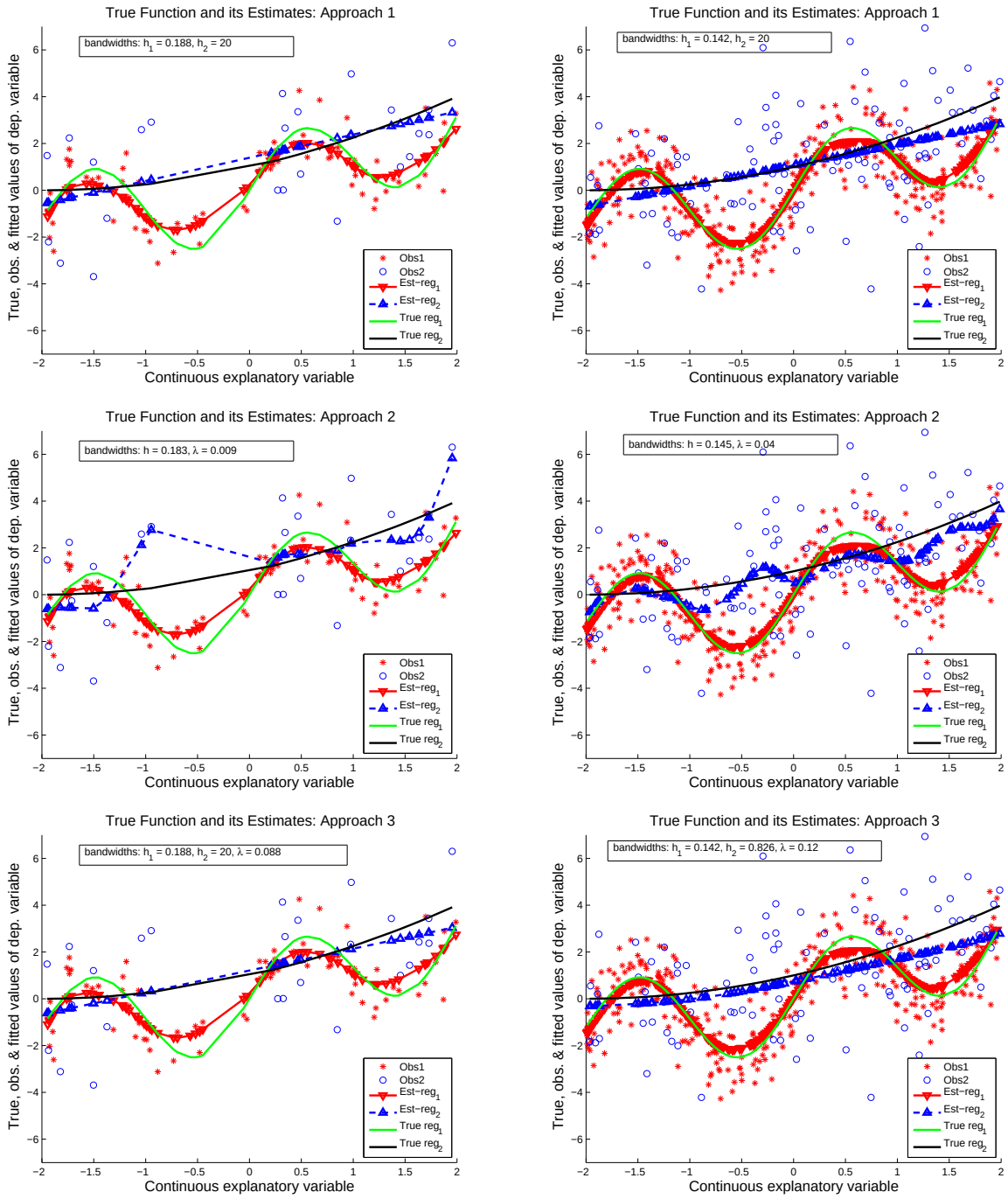

Figure 5: *Example 5: Left panel,* $n = 100$ *, and right panel,* $n = 400$*.From top to bottom: Approach1 (no-smoothing the discrete variable), Approach2 (simple- smoothing), Approach3 (complete-smoothing).*

This issue appears to have been overlooked in previous studies, in particular in the context of local linear fitting.

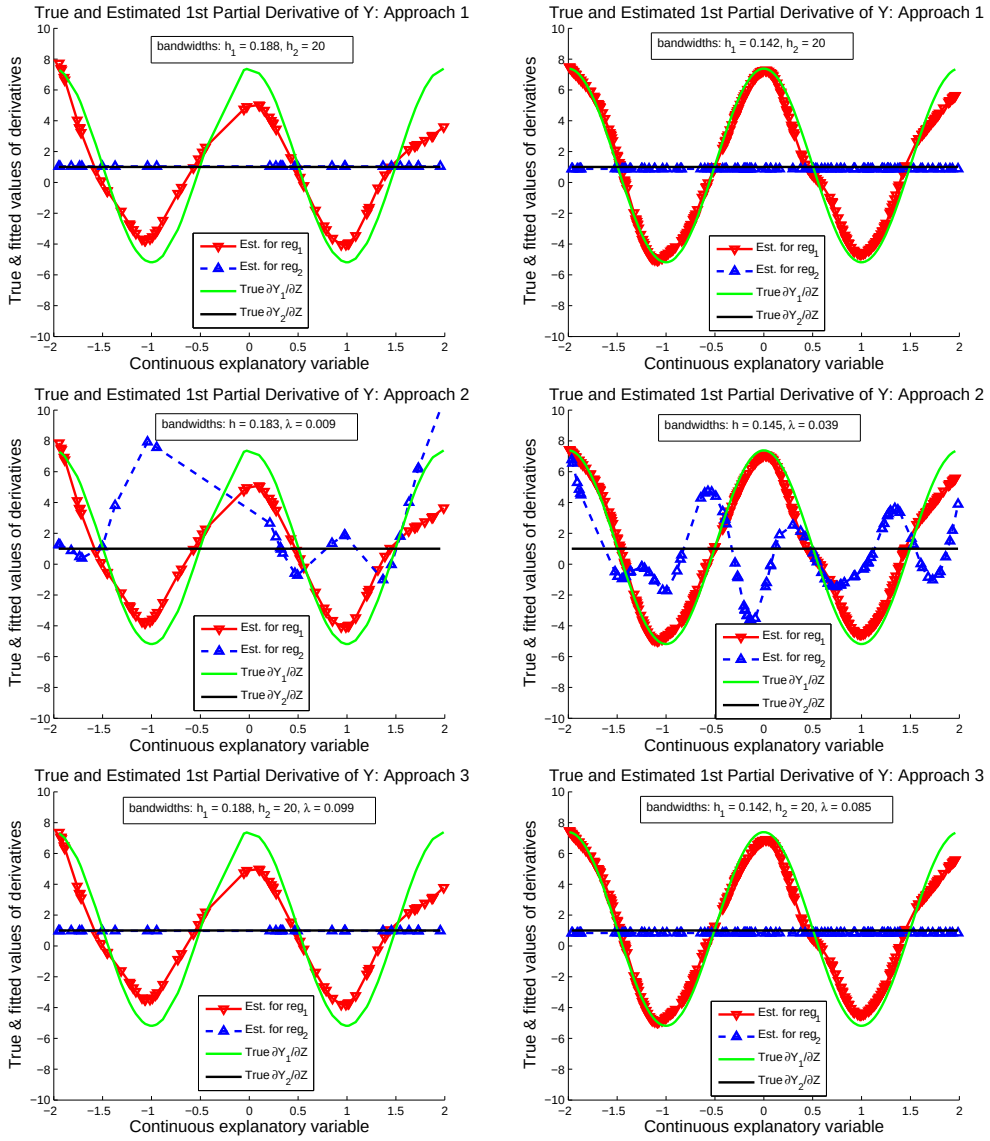

Figure 6: *Derivative Estimates for Example 1: Left panel,* $n = 100$ *and right panel,* $n = 400$*. From top to bottom: Approach1 (no-smoothing the discrete variable), Approach2 (simple-smoothing), Approach3 (complete-smoothing).*

We have also seen that if the difference between the DGP of the two groups is relatively small, e.g., as in Example 3, the simple-smoothing of the discrete variable is working well. But even there, as expected, the extended Racine-Li approach provides significantly better fit. It is thus clear that, even for fairly small samples, there are cases where fully separate estimation, without smoothing the discrete variable, performs significantly better than the simple-smoothing with common bandwidths for the continuous variables across groups. In all the cases we considered, the complete-smoothing approach behaves better,

but at a computational cost. We also mentioned that the problem of derivatives estimations is contaminated by the same phenomenon, with all the consequences this may lead to.

5 Empirical Illustration

The goal of this section is to make an illustration of the phenomenon we discussed above for the context of a real data. To do so, we will use a data set from a study of Kumar and Russell (2002), about patterns of convergence or divergence in economic growth in the world.[5] We choose this data and the context because the topic of economic growth has remained interesting for a wide audience for centuries that past and, perhaps, the centuries that follow, and so we hope it would be interesting to a general audience.

This data consists of observations on 57 countries in the world, containing such variables as GDP, labour and capital of each country in 1965 and in 1990, and was originally extracted from the Penn World Tables. We will use this data to estimate regression relationship between the growth in GDP per capita (between 1965 and 1990) of countries in the world (used as dependent variable here) and the initial levels of GDP per capita of these countries (used as the continuous explanatory variable). Such a regression and many of its variations are often performed in empirical economic growth studies on convergence.

Often, an hypothesis of interest in such studies is that the growth rates of poorer countries are, on average, higher than those of the richer countries, and so the poorer countries eventually must catch-up with or converge to the levels of GDP per capita of the richer countries. This is often referred to as the (unconditional) "beta-convergence" phenomenon. Earlier works on this issue employed parametric regressions and some studies found that the slope coefficient (the "beta") in such regressions, is negative and significantly different from zero, thus supporting the "beta-convergence" hypothesis. Yet, other studies found that the "beta" is insignificantly different from zero (i.e., no convergence) or even positive ("beta-divergence") and significantly different from zero for different samples or for distinct groups of countries within a sample or when additional explanatory variables are accounted for.[6] Clearly, any of such results might also depend on the parametric assumptions on the regression relationship and so this is where using non-parametric regression methods may give some useful insights. Below we will use the local linear least squares estimator (LLLSE), with the three approaches discussed above, to see whether they suggest the same or similar

[5]This data set (or its extended version) was also used in many other applications, e.g., in Henderson and Russell (2005), Simar and Zelenyuk (2006), Henderson and Zelenyuk (2007), Badunenko et al. (2008), etc.

[6]For a recent detailed review of this topic, e.g., see Maasoumi et al. (2007), Weil (2008) and references cited therein.

stories for the following regression relationship,

$$y_i = m(z_i, z_i^d) + \varepsilon_i,\ i = 1, \ldots, n$$

where y_i is growth in GDP per capita of country i between 1965 and 1990, z_i is the natural log of GDP per capita of country i in 1965, while z_i^d is a discrete variable (defined below) and ε_i is statistical noise, for which we assume that $E(\varepsilon_i|z_i, z_i^d) = 0$ and $V(\varepsilon_i|z_i, z_i^d) < \infty$ for all i.

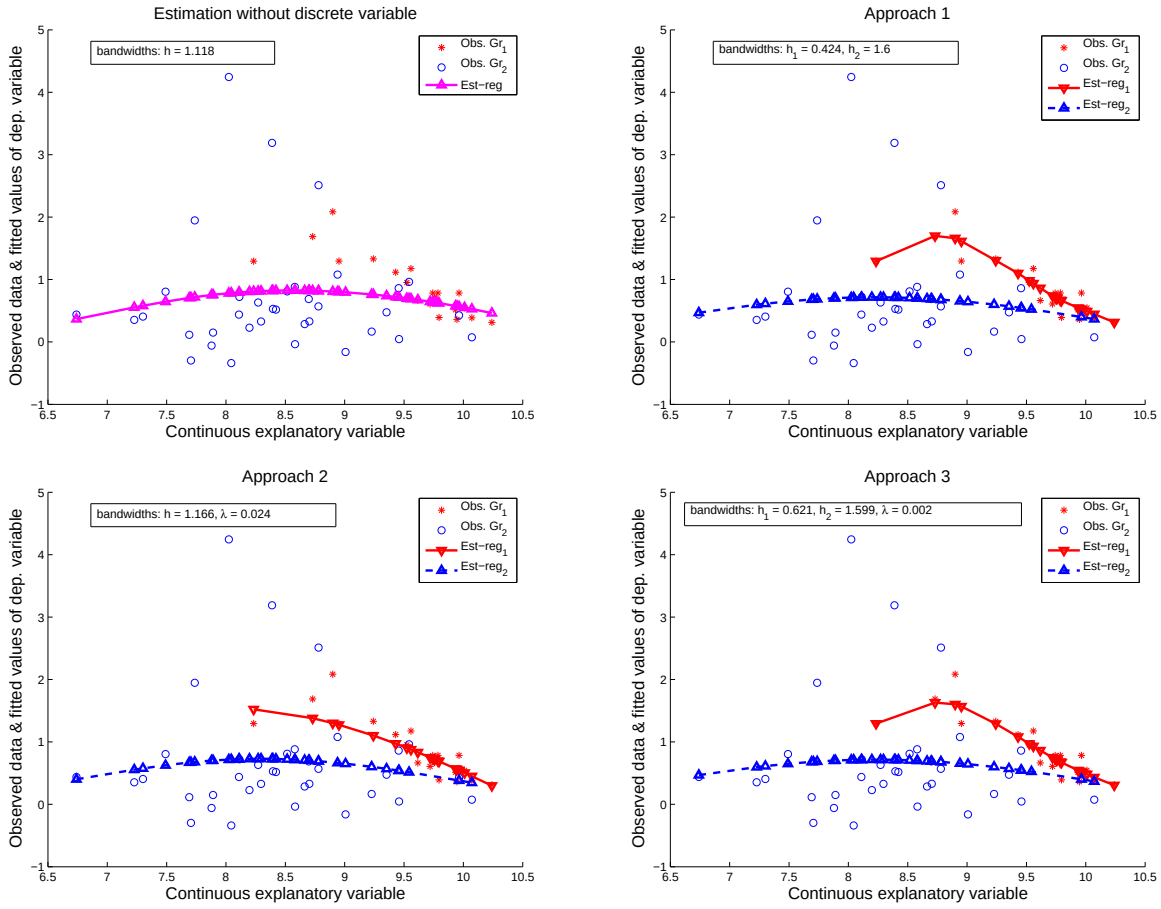

Figure 7: *Illustration with GDP data. From left to right and top to bottom, Panel (a) Approach 0 (only one group of data), Panel (b): Approach1 (no-smoothing the discrete variable), Panel (c): Approach 2 (simple-smoothing), Panel (d): Approach 3 (complete-smoothing).*

The result of the estimations are shown in Figure 7 including some information on the resulting bandwidths obtained by the different approaches. As a starting point, in panel (a) of Figure 7 we present results of LLLSE estimation without the discrete variable (where

h is estimated with LSCV). From this figure one may get a feeling that there are might be different groups within the sample that have different regression relationships (in terms of intercept or slope or both). Indeed, in various studies in the context of cross-countries analyses, researchers often distinguish various groups of countries, allowing them to have different regression relationships. An objective grouping criterion often used in practice, for example, is an indicator whether a country is OECD member or not (e.g., Racine et al. (2006), Simar and Zelenyuk (2006), Maasoumi et al. (2007), Henderson and Zelenyuk (2007), etc.), and so we also use it as our discrete variable, z_i^d, that has value 1 if country i was a member of OECD in the year 1965 and zero otherwise.[7]

In panel (b) of Figure 7 we present results of LLLSE estimation for Approach 1 (i.e., separate estimation for each group, with h estimated via LSCV for each group separately), and one can see that the estimated relationships for the two groups are very different not only in the intercept but also in the slope. Specifically, note that the relationship for the larger (non-OECD) group is virtually flat, with very slight inverted-U-shape curvature. On the other hand, note that the relationship for the smaller (OECD) group has a more pronounced inverted-U-shape (or rather "inverted hockey-stick" shape) curvature. Note that such curvature may suggest an important economic implication: it hints that OECD countries with very low initial GDP per capita are expected to have higher growth rates in GDP per capita than those with very high initial GDP per capita, yet the highest rates are expected to be not at the lowest level of GDP per capita but somewhat larger.

On the other hand, panel (c) of Figure 7 presents results of LLLSE estimation for Approach 2 (i.e., original Racine-Li approach with OECD variable, where h and λ are estimated jointly via LSCV for the entire sample). One can see that the estimated relationships here are also very different between the two groups, but also somewhat different from the story suggested by the Approach 1 in panel (b). Specifically, note that the relationship for the larger (non-OECD) group has slightly more pronounced inverted-U-shape curvature, although it still remains relatively flat. On the other hand, note that the relationship for the smaller (OECD) group has much less curvature than was observed from Approach 1 in panel (b), which is not an inverted-U-shape at all. That is, Approach 2 suggests that for OECD countries there is almost linear and negative relationship between the growth in GDP per capita and the initial level of GDP per capita. In other words, with the simple-smoothing approach we get some under-smoothing for the larger group and over-smoothing for the smaller group relative to the Approach 1 (separate estimation). This is similar to what we have observed

[7]Clearly, in a detailed analysis, one may want to condition for many other potentially important explanatory variables, yet we will limit our illustration to the case of one continuous and one discrete explanatory variable for the sake of ease of graphical representation of the phenomenon.

for simulated data sets, except that now we do not know how the true relationships look like.

Some additional insight is provided by the Approach 3, i.e., the complete-smoothing method, where we allow for each group identified by the discrete variable to have its own bandwidth but also smooth the discrete variable and so use the full sample in one estimation. Panel (d) of Figure 7 visualizes the results of the estimation from Approach 3 and one can see that it gives almost identical results to those from Approach 1. Again, this is similar to what we have observed for simulated data sets, yet now we do not know what the true relationship is. Indeed, with such small samples, it could be that the left-most observation in group 1 is kind of 'accidental' and so omitting it for Approach 1 and 3 may give results very similar to Approach 2. However, it also could be the case that there are other data points not available in our sample that are similar to this left-most observation in group 1 and including them would make the inverted U-shape curvature even more pronounced. Since we do not know the true relationship, unlike in the simulated samples, it is hard to judge which of these arguments is likely to be right or wrong and we do not do so, but only point out that Approach 1 and Approach 3 gave almost identical results, which are different from Approach 2. Since the Approach 3 encompasses the other two approaches, taking the best features from each, and that our simulations suggested that Approach 3 was never worse than the other two, and sometimes better than at least one of them, Approach 3 appears to be the safest approach for a practitioner to trust in this context and, perhaps in general, whenever it is computationally feasible.

Finally, it might be worth emphasizing again that in this section we had not intended to resolve the puzzles of economic growth across countries as such study would require larger data set, more variables, etc. Our goal was just to give a concise and vivid illustration of the phenomenon we discussed above and, in particular, to compare the three approaches, not only for simulated data sets, but also for a real data, for a context that appears to be interesting for a wide audience.

6 Conclusion

In this article we pointed out and illustrated that the reduction in variance or the efficiency gain due to smoothing of the discrete regressors with common bandwidth for the continuous variables across groups, as is frequently done in applied studies so far, can be well outweighed by the substantial bias introduced due to this smoothing, both for small and for relatively large samples. For such cases, even fully separate estimation for each group, if feasible, might be preferred. We have shown that the extended smoothing technique (allowing different

bandwidths for the continuous variables in each group) overcomes this difficulty and thus is much more safe: it outperforms the two other approaches, but at an additional computational cost.

In general, whether it is better to smooth or not to smooth the discrete variable or whether "the bias beats the variance", or not, essentially depends on the degree of difference of the DGPs in the different groups: curvatures of the regression relationship, variation in the error term for each group, variation in the continuous regressors, size or proportion of one group relative to another in the sample, etc... The safe approach is indeed an extension of the "default" Racine and Li method, but at a cost of additional computational complexity that could quickly become problematic either if the number of continuous variables increases (and we use a multiplicative kernel) or when the number of categories determined by the discrete variables is too large or both. The numerical burden is linked to the determination of optimal bandwidths by solving nonlinear optimization problem in very high dimensions. This burden can be reduced if one is willing to make additional assumptions on common "degree of smoothness" of the regression function in some subgroups, and so have common h for these groups, but at the same time be flexible for other groups. Clearly, such decisions impose additional structure on the model, which is more flexible than in the parametric approaches, yet it is more restrictive than in a fully non-parametric estimation, and so may need some out-of-sample information and justifications.

In this respect, we can also say that by using the default or simple smoothing of discrete variables in non-parametric regression one automatically (or implicitly) imposes the assumption of similar degree of smoothness of the regression relationships in all the groups of the sample identified by the discrete variables, which might be far from reality. As we illustrated with several examples, such restriction can significantly deteriorate estimation results, increasing bias in the estimates of the true regression relationship. This same problem can also substantially or even radically distort estimates of derivatives and the related estimates of marginal effects, elasticities, etc. that are used to draw policy implications. For example, with simple smoothing one could obtain radically varying and even changing the sign estimates of derivatives while their true values (and the values estimated via complete-smoothing or separate estimation) could in fact be constant or vice versa. This means that research conclusions, policy implications and consequently the real policy decisions based on such estimates can be misleading or wrong. Therefore, this implicit assumption about common h is especially important to realize and explicitly acknowledge as 'an extra price to pay' for using the simple smoothing, instead of fully separate or complete-smoothing approaches.

It is also important to recognize that even if from a theoretical point of view, the extended Racine and Li smoothing of the discrete variable is preferable and offers a suitable solution

to the problem, it is still an open question in practice, for a real data set, if we have to smooth or not to smooth the discrete variables, particularly when the computational cost of the extended method is prohibitive. Further theoretical work is thus needed to develop and justify a method (a statistical test, a rule of thumb, ...) that would help justifying a decision whether to smooth or not to smooth over some or all discrete variables. The issue of relevance of some categorical predictors in nonparametric regressions has been analyzed in Racine et al. (2006). They consider testing the hypothesis $\lambda_\ell = 1$, but to the best of our knowledge nothing has been done, at the other extreme of the scale ($\lambda_\ell = 0$), including the issue of common bandwidths for the continuous variables. Another extension of our work would be to investigate whether the complete-smoothing we proposed in this paper can also improve performance of various tests that employ discrete smoothing (e.g., see Racine et al. (2006), Hsiao et al. (2007), etc.), as it would be a natural consequence of what we find in the present work. It was the purpose of this paper to warn the practitioners of the caveats of the simple-smoothing method, to suggest a safer procedure, when it is doable in practice, and to call for complementary theoretical efforts to address these imperative issues.

References

[1] Aitchison, J. and C.G.G. Aitken (1976), Multivariate binary discrimination by the kernel method, *Biometrika*, 63, 3, 413–420.

[2] Badunenko, O. Henderson, D. and V. Zelenyuk (2008), Technological change and transition: relative contributions to worldwide growth during the 1990s, *Oxford Bulletin of Economics and Statistics*, 70:4, 461–492.

[3] Cleveland, W.S. (1979), Robust locally weighted regression and smoothing scatterplots, *Journal of American Statistical Association*, 74, 829–836.

[4] Cleveland, W.S. and S.J. Delvin (1988), Locally-weighted regression: an approach to regression analysis by local fitting, *Journal of American Statistical Association*, 83, 579–610.

[5] Eren, O. and D. Henderson (2008), The impact of homework on student achievement, *Econometrics Journal*, 11, 326–348.

[6] Fan, J. (1992), Design-adaptative nonparametric regression, *Journal of American Statistical Association*, 87, 998–1004.

[7] Fan, J. (1993), Local linear regression smoothers and their minimax efficiency, *Annals of Statistics*, 21, 196–216.

[8] Fan J. and I. Gijbels (1992), Variable bandwidth and local linear regression smoothers, *Annals of Statistics*, 20, 2008–2036.

[9] Fan J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall.

[10] Gozalo, P. and O. Linton (2000), Local nonlinear least squares: Using parametric information in nonparametric regression, *Journal of Econometrics*, 99, 63–106.

[11] Hall, P., Li, Q. and J. Racine (2007), Nonparametric estimation of regression functions in the presence of irrelevant regressors, *The Review of Economics and Statistics*, 89, 4, 784–789.

[12] Hartarska, V., C.F. Parmeter and D. Nadolynak (2010), Economies of scope of lending and mobilizing deposits in Microfinance institutions: A semiparametric Analysis, *American Journal of Agricultural Economics*, 93(2), 389–398.

[13] Henderson, D. (2010), A test for multimodality of regression derivatives with an application to nonparametric growth regressions, *Journal of Applied Econometrics*, 25, 458–480.

[14] Henderson, D.J. and Russell, R.R. (2005), Human capital and convergence: a production-frontier approach, *International Economic Review*, Vol. 46, 1167–1205.

[15] Henderson, D.J. and Zelenyuk, V. (2007), Testing for (efficiency) catching-up, *Southern Economic Journal*, Vol. 73, 1003–1019.

[16] Hsiao, C, Q. Li and J. Racine (2007), A consistent model specification test with mixed discrete and continuous data, *Journal of Econometrics*, 140, 802–826.

[17] Kumar, S., and R.R. Russell (2002), Technological change, technological catch-up, and capital deepening: Relative contributions to growth and convergence, *American Economic Review*, 92, 527–48.

[18] Li, Q. and J. Racine (2004), Cross-validated local linear nonparametric regression, *Statistica Sinica*, 14, 485–512.

[19] Li, Q. and J. Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[20] Maasoumi, E., J. Racine, and T. Stengos (2007), Growth and convergence: A profile of distribution dynamics and mobility, *Journal of Econometrics*, 136(2), 483–508.

[21] Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press.

[22] Park, B., Simar, L. and V. Zelenyuk (2010), Local Maximum Likelihood Methods with Categorical Variables. Discussion paper 1052, Institut de Statistique, UCL.

[23] Parmeter, C.F., D. Henderson and S.C. Kumbhakar (2007), Nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics*, 22, 695–699.

[24] Racine, J., Hart, J. and Q. Li (2006), Testing the significance of categorical predictor variables in nonparametric regression models, *Econometric Review*, 25(4), 523–544.

[25] Racine, J. and Q. Li (2004), Nonparametric estimation of regression functions with both categorical and continuous data, *Journal of Econometrics*, 119, 99–130.

[26] Ruppert, D. and M.P. Wand (1994), Multivariate weighted least squares regression, *Annals of Statistics*, 22, 1346–1370.

[27] Simar, L. and V. Zelenyuk (2006), On Testing Equality of Two Distribution Functions of Efficiency Score Estimated via DEA, *Econometric Reviews*, 25(4), 497-522.

[28] Stengos, T. and E. Zacharias (2006), Intertemporal pricing and price discrimination: a semiparametric hedonic analysis of the personal computer market, *Journal of Applied Econometrics*, 21(3), 371–386.

[29] Stone, C. J. (1977), Consistent nonparametric regression, *Annals of Statistics*, 5, 595–645.

[30] Titterington, D. M. (1980), A comparative study of kernel-based density estimates for categorical data, *Technometrics*, Vol 22, 2, 259–268.

[31] Walls, W. (2009), Screen wars, star wars, and sequels, *Empirical Economics*, 37(2), 447–461.

[32] Wang, M.C. and J. Van Ryzin (1981), A class of smooth estimators for discrete distributions, *Biometrika*, 68, 301–309.

[33] Weil, D.N. (2008), *Economic Growth*, 2nd ed. Addison Wesley.

Table 1: Monte-Carlo Results for Example 1, over 200 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3
$\overline{AMSE}$ all	0,3904	0,5789	0,3578	0,2030	0,3861	0,2206	0,1083	0,2180	0,1156	0,0636	0,1307	0,0649
std_{MC} all	0,0150	0,0175	0,0115	0,0062	0,0095	0,0066	0,0029	0,0050	0,0033	0,0019	0,0028	0,0018
$\overline{AMSE}$ gr1	0,3065	0,3641	0,2888	0,1632	0,2090	0,1806	0,0923	0,1130	0,1004	0,0544	0,0659	0,0585
std_{MC} gr1	0,0166	0,0189	0,0104	0,0047	0,0068	0,0058	0,0025	0,0034	0,0027	0,0014	0,0020	0,0017
$\overline{AMSE}$ gr2	0,7200	1,2890	0,5888	0,3391	0,9326	0,3473	0,1558	0,5430	0,1639	0,0918	0,3249	0,0845
std_{MC} gr2	0,0515	0,0564	0,0341	0,0244	0,0402	0,0208	0,0091	0,0179	0,0097	0,0063	0,0105	0,0053
median $\hat{h}_1$	0,2159	0,2056	0,2159	0,1874	0,1897	0,1879	0,1632	0,1566	0,1632	0,1421	0,1395	0,1421
median $\hat{h}_2$	20,000	0,2056	20,000	20,000	0,1897	20,000	20,000	0,1566	20,000	20,000	0,1395	20,000
median $\hat{\lambda}$	-	0,1095	0,0986	-	0,0755	0,0533	-	0,0535	0,0195	-	0,0357	0,0110

Table 2: Monte-Carlo Results for Example 2, over 200 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3
$\overline{AMSE}$ all	0,2949	0,3588	0,2245	0,1526	0,2288	0,1299	0,0816	0,1295	0,0732	0,0466	0,0740	0,0429
std_{MC} all	0,0128	0,0137	0,0098	0,0064	0,0082	0,0051	0,0035	0,0040	0,0027	0,0020	0,0022	0,0017
$\overline{AMSE}$ gr1	0,1381	0,1766	0,1490	0,0764	0,1006	0,0921	0,0403	0,0570	0,0546	0,0250	0,0307	0,0296
std_{MC} gr1	0,0063	0,0087	0,0072	0,0033	0,0048	0,0040	0,0015	0,0024	0,0021	0,0009	0,0012	0,0012
$\overline{AMSE}$ gr2	0,8370	0,9307	0,4510	0,3994	0,6190	0,2434	0,2049	0,3524	0,1291	0,1121	0,2031	0,0830
std_{MC} gr2	0,0572	0,0524	0,0329	0,0282	0,0337	0,0162	0,0133	0,0156	0,0082	0,0076	0,0084	0,0050
median $\hat{h}_1$	0,3631	0,4514	0,3672	0,3250	0,3976	0,3758	0,2758	0,3589	0,3263	0,2292	0,3095	0,2846
median $\hat{h}_2$	20,000	0,4514	20,000	20,000	0,3976	20,000	20,000	0,3589	20,000	20,000	0,3095	20,000
median $\hat{\lambda}$	-	0,1308	0,2179	-	0,0940	0,1822	-	0,0570	0,1511	-	0,0413	0,1065

Table 3: Monte-Carlo Results for Example 3, over 200 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3
$\overline{AMSE}$ all	0,2445	0,2142	0,1540	0,1264	0,1325	0,0916	0,0633	0,0709	0,0493	0,0385	0,0474	0,0299
std_{MC} all	0,0116	0,0139	0,0084	0,0055	0,0083	0,0044	0,0026	0,0034	0,0020	0,0017	0,0018	0,0011
$\overline{AMSE}$ gr1	0,1083	0,1161	0,1119	0,0602	0,0766	0,0694	0,0320	0,0461	0,0411	0,0207	0,0271	0,0231
std_{MC} gr1	0,0055	0,0053	0,0054	0,0030	0,0033	0,0032	0,0015	0,0020	0,0017	0,0009	0,0012	0,0010
$\overline{AMSE}$ gr2	0,7200	0,5371	0,2801	0,3391	0,3063	0,1601	0,1558	0,1447	0,0739	0,0920	0,1079	0,0502
std_{MC} gr2	0,0515	0,0580	0,0285	0,0244	0,0344	0,0144	0,0091	0,0123	0,0063	0,0063	0,0066	0,0034
median $\hat{h}_1$	0,8029	1,0263	0,8322	0,6914	0,8789	0,6409	0,5427	0,6541	0,4950	0,4214	0,5381	0,3956
median $\hat{h}_2$	20,000	1,0263	20,000	20,000	08789	20,000	20,000	0,6541	20,000	20,000	0,5381	4,0017
median $\hat{\lambda}$	-	0,3859	0,4190	-	0,3329	0,3652	-	0,2879	0,3124	-	0,2503	0,3000

Table 4: Monte-Carlo Results for Example 4, over 200 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3
$\overline{AMSE}$ all	0,2917	0,2815	0,2132	0,1598	0,1810	0,1382	0,0873	0,0948	0,0714	0,0508	0,0621	0,0447
std_{MC} all	0,0131	0,0144	0,0081	0,0071	0,0084	0,0049	0,0038	0,0038	0,0027	0,0021	0,0022	0,0016
$\overline{AMSE}$ gr1	0,1156	0,1283	0,1256	0,0623	0,0862	0,0774	0,0326	0,0473	0,0391	0,0211	0,0274	0,0229
std_{MC} gr1	0,0058	0,0058	0,0057	0,0031	0,0038	0,0036	0,0015	0,0021	0,0017	0,0009	0,0012	0,0009
$\overline{AMSE}$ gr2	0,8943	0,7655	0,4830	0,4709	0,4723	0,3246	0,2497	0,2402	0,1716	0,1397	0,1664	0,1103
std_{MC} gr2	0,0587	0,0577	0,0275	0,0314	0,0339	0,0171	0,0150	0,0147	0,0095	0,0078	0,0084	0,0057
median $\hat{h}_1$	0,7071	0,8612	0,7277	0,6194	0,7351	0,5277	0,4995	0,6556	0,4437	0,3879	0,5296	0,3775
median $\hat{h}_2$	3,2648	0,8612	20,000	20,000	0,7351	1,7982	2,1567	0,6556	4,0005	1,7787	0,5296	1,3333
median $\hat{\lambda}$	-	0,1864	0,2255	-	0,1374	0,1406	-	0,0890	0,0770	-	0,0698	0,0514

Table 5: Monte-Carlo Results for Example 5, over 200 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3	Appr 1	Appr 2	Appr 3
$\overline{AMSE}$ all	0,4381	0,5716	0,3909	0,2429	0,3856	0,2556	0,1383	0,2178	0,1375	0,0812	0,1307	0,0853
std_{MC} all	0,0159	0,0166	0,0119	0,0074	0,0095	0,0084	0,0040	0,0050	0,0038	0,0022	0,0028	0,0021
$\overline{AMSE}$ gr1	0,3066	0,3533	0,2872	0,1632	0,2098	0,1814	0,0924	0,1136	0,1022	0,0544	0,0662	0,0601
std_{MC} gr1	0,0165	0,0172	0,0101	0,0047	0,0068	0,0057	0,0025	0,0034	0,0027	0,0014	0,0020	0,0017
$\overline{AMSE}$ gr2	0,9175	1,2926	0,7276	0,5025	0,9282	0,4829	0,2754	0,5403	0,2456	0,1616	0,3239	0,1609
std_{MC} gr2	0,0585	0,0563	0,0395	0,0312	0,0402	0,0308	0,0149	0,0178	0,0143	0,0078	0,0105	0,0071
median $\hat{h}_1$	0,2159	0,2051	0,2159	0,1874	0,1898	0,1879	0,1632	0,1566	0,1632	0,1421	0,1392	0,1421
median $\hat{h}_2$	2,2241	0,2051	2,4473	3,7253	0,1898	1,7338	1,3449	0,1566	1,3761	1,0003	0,1392	0,9094
median $\hat{\lambda}$	-	0,1111	0,0972	-	0,0766	0,0572	-	0,0540	0,0270	-	0,0360	0,0192